# Addressing key big data analytics challenges in healthcare domain using artificial intelligence

Synopsis for the Doctoral Degree in Computer Science and Engineering

Submitted By

**Chetan Ganpat Malavade**

Under the Guidance of

**Dr. Megha Venkatesh Jonnalagedda**
Associate Professor
Department of Information Technology
Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded

Submitted to
Research and Recognition Committee

**SWAMI RAMANAND TEERTH MARATHWADA UNIVERSITY**
**NANDED – 431606, Maharashtra State, India**

# Contents

1. Introduction ---------- 3

2. Literature Review ---------- 7

3. Objectives ---------- 9

3.1 Expected Outcomes ---------- 9

4. Research methodology ---------- 10

5. Importance of study (applicability to society) ---------- 12

6. Proposed work plan ---------- 13

7. References ---------- 14

# 1. Introduction

Currently the entire world is transitioning to Industry 4.0 (fourth industrial revolution) where digital technologies are playing a vital role. This is especially true for the healthcare domain where Internet of Things, Big Data Analytics, Blockchain, Cloud Computing and Artificial Intelligence/ Machine Learning enabled technologies are revolutionizing the entire healthcare domain ecosystem end to end.

With the increase in the popularity and easy accessibility of the internet, the consumers' digital footprints over the internet have accelerated these days - especially post current pandemic. It is believed that over 2.5 quintillion bytes of data is created every day, and this number is in increasing order. Interestingly, it is said that over 90% of these data are collected within just the past 2 years. Internet queries, social media posts, changed communication mechanisms and changed service industries including healthcare industry are greatest contributors to this big data. This data can be of structured, unstructured, and semi-structured type. Different varieties of data include the text, audio, video, log files, sensor data etc. in petabytes and beyond.

Big data analytics is gaining great attention these days. Both industry and academy researchers are working on numerous problems in this domain where a humongous amount of data is generated daily. Identifying correlation, understanding trends in these massive datasets surely would be beneficial across health care, social media, agriculture, finance, and education domains. Noise, incompleteness, and inconsistency in this data collected is going to cause uncertainty in this analysis. Advanced analytical techniques, efficient review mechanisms, adaptive strategies are to be crafted to tackle this humongous data efficiently.

Big data analytics has the potential to transform the way healthcare providers use sophisticated technologies to gain insight from their clinical and other data repositories and make informed decisions. As big data analytics becomes more mainstream, issues such as guaranteeing privacy, safeguarding security, establishing standards for governance, and continually improving the tools and technologies are going to receive attention. Efficient implementation of big data analytics for the healthcare domain is significant. Accomplishment of big data analytics in healthcare surely is going to have a lot many assertive and life-protecting end results.

In comparison to traditional data techniques and platforms, artificial intelligence techniques (including machine learning, natural language processing, and computational intelligence) provide more accurate, faster, and scalable results in big data analytics. Data privacy and confidentiality are the most important challenges in healthcare big data analytics. Security is the major hindrance in the broad integration of big data analytics, particularly when dealing with patient's health data.

While big data analytics using AI holds a lot of promise, a wide range of challenges are introduced when such techniques are subjected to uncertainty. Typically, the data collected from sensors, social media, financial records, etc. is inherently uncertain due to noise, incompleteness, and inconsistency. The analysis of such inaccurate data requires advanced analytical techniques for efficiently reviewing and/ or predicting future courses of action with high precision and advanced decision-making strategies. As the amount, variety, and speed of data increases, so does the uncertainty inherent within, leading to a lack of confidence in the resulting analytics process and decisions made thereof. Key healthcare big data analytics challenges enlisted as follows:

**Data security and privacy:** Data Security, privacy, integrity, and lack of transparency are becoming real challenges in healthcare as more and more data is digitized. Most of the people are not willing to share their personal data with a fear of security breach. If there is assurance for data security, then the problem can be managed. Strong technological hardware and software level security precautions and measures should be implemented to prevent the hacking and interpreting malicious code. There are many chances in IoT that the data security and privacy may get compromised as the data being sent and received among various IoT-enabled devices lack various standard data transmission protocols, and ambiguity may arise regarding data ownership and its regulation. These factors make data prone to hackers and crackers, and because of which the personal health information (PHI) may get compromised and both patient and doctor may have to suffer from the loss due to the same.

**Data accuracy:** Healthcare data inconsistencies, lack of precision and varied data along different timelines may lead to data inaccuracy. Some patients may hide personal details about their lifestyle during filling up forms in hospitals/ clinics or during face-to-face interactions with doctors. Sometimes it

may happen that some of the fields produce wrong results. If analysis is done on the empty or wrong fields of data, then it may or may not get processed. In both cases they produce wrong results.

**Real time integration:**Real time integration of the huge amount of data in multiple forms (structured, unstructured, semi-structure) from a variety of sources (novel and traditional sources) is also going to be one of the key pain areas here. Healthcare data does not remain the same always. Most of the data elements require frequent updates to stay relevant. Understanding the sensitivity of big data, or how often and to what extent it should change, is a challenge for all healthcare providers who don't consistently monitor their datasets. These dataset updates can happen manually or in an automated way. This complete process should be seamless (without any downtimes, no damage to data quality and integrity as such).

**Data storage:** Healthcare professionals are not concerned about medical data storage however for IT teams designing and developing healthcare applications this information is very crucial as it may raise security, performance risks. Healthcare data will continue to grow and so will be the storage costs. On premise data centers are used however this approach can be expensive to scale, difficult to maintain, and prone to producing data silos across various departments within healthcare organization. On the other hand, cloud storage offers nimble disaster recovery, lower up-front costs, and easy expansion. Healthcare providers can go for a hybrid approach for their data storage programs. While developing a hybrid infrastructure, providers must ensure that disparate systems can communicate and share data with other segments of the organization if required.

**Interoperability:** This challenge may arise in healthcare organizations when they wish to actively participate in a larger ecosystem. All stakeholders in the ecosystem must follow a digitized approach. There might still be semantic challenges as applications which are interacting with each other are likely to use different terminology to describe the same thing.

**Unstructured data:** Healthcare Big Data Analytics process can collect data from various sources which would be unstructured data like prescriptions, blogs, tweets, healthcare records, status updates, doctor comments etc. It is necessary to generate right metadata from this unstructured data and transform it into a structured format by applying intelligent processing techniques to filter meaningful insights.

**Human errors:** Electronic health record data is influenced by the person who entered the patient's data, which can lead to missing data, incorrect data, misunderstanding or wrong interpretation of the original data.

Proposed research work will address some of these key challenges and attempt to minimize damages occurring due to current limitations in the healthcare domain. Novel approaches would be proposed, and effective system design would be recommended by detailed study of currently open challenges and addressing those in most efficient ways.

## 2. Literature Review

Giuseppe Aceto et al. [1] provide insights regarding main technologies and paradigms in relation to Healthcare 4.0. Paper talks about Healthcare 4.0 application scenarios and novel cross-disciplinary challenges. Shuva Paul et al. [2] analyze various healthcare domain applications, currently available state-of-the-art technologies and review solutions proposed by innovative start-ups to address challenges with regards to the Industry 4.0 paradigm. These challenges and proposed future work would be thoroughly analyzed to come up with novel solutions.

Siva Sankara Reddy et al. [3] describe various steps involved in the Big Data Analytics process and discuss its advantages and challenges which show impact on healthcare organization. It also presents how to analyze the facts by using big data analytics to improve the healthcare in the world overall. Fatima Lalmi and LaadjalAdala [4] in their paper discuss various challenges and opportunities in big data analytics mainly associated with the healthcare domain. They have also suggested five phases of gradual transformation into a healthcare system based on big data analytics. Healthcare reforming technological solutions would be proposed as a part of planned research work enhancing the level of healthcare delivery as well as supporting decision-making.

Rony Chowdhury Ripan et al. [5] propose a data-driven prediction model for anomaly detection in the healthcare domain to effectively predict heart disease using unsupervised K-means clustering algorithms. Krishnan Naidoo et al. [6] evaluate different machine learning models and propose an unsupervised framework to identify anomalies using a Generative Adversarial Network model. Menaa Nawaz and Jameel Ahmed [7] implemented an IoT based healthcare system using bio-medical sensors. This paper provides detailed analysis of cloud data acquired through biomedical sensors using signal analysis techniques for anomaly detection and time series data classification done for the disease prognosis in real-time. These models can be further enhanced to achieve better accuracy on recent datasets even for different diseases which are impacting masses.

Anonymization or removal of patient identifiable information, though the most conventional way, is the first important step to adhere to the regulations and incorporate privacy concerns. Iyiola E. Olatunji et al. [8] provides a comprehensive review of anonymization models and techniques applicable for relational and graph-based health care data. Planned research work would aim to preserve privacy, maintain utility

without loss of association for both numerical as well as categorical MSAs.Jayapradha. J and Prakash. M [9] focuses on enabling the anonymization of multi-valued records of an individual. They implemented vertical partitioning and applied k-anonymity for quasi-identifier bucket (QIAB) and (k, l) diversity for individual multi-sensitive attribute bucket (IMDB). These results were validated only for static data. Further exploration would be done on streaming data/ unstructured data using the same approach for recent healthcare datasets. Veena Gadad et al. [10] propose an effective privacy preserving algorithm mainly to address multiple sensitive attributes to minimize information loss. All such existing algorithms could be compared/ evaluated on factors like time, data utility and information losses occurring.

Jahoon Koo et al. [11] identify threats and security issues that occur in the life cycle of big data. They divide a big data life cycle into five phases (i.e., collection, storage, analytics, utilization, and destruction), and define the security taxonomy of the big data life cycle based on the identified threats and security issues. Proposed research work will attempt to have security architecture in-line with suggested security taxonomy.Arshia Rehman et al. [12] present emerging landscape of big data and analytical techniques in the five subdisciplines of healthcare namely medical image analysis andimaging informatics, bioinformatics, clinical informatics, public health informatics and medical signal analytics.Available methods, tools, technologies and architectures are discussed at length along with futuristic challenges.

# 3. Objectives

Proposed research work objectives are enlisted below.

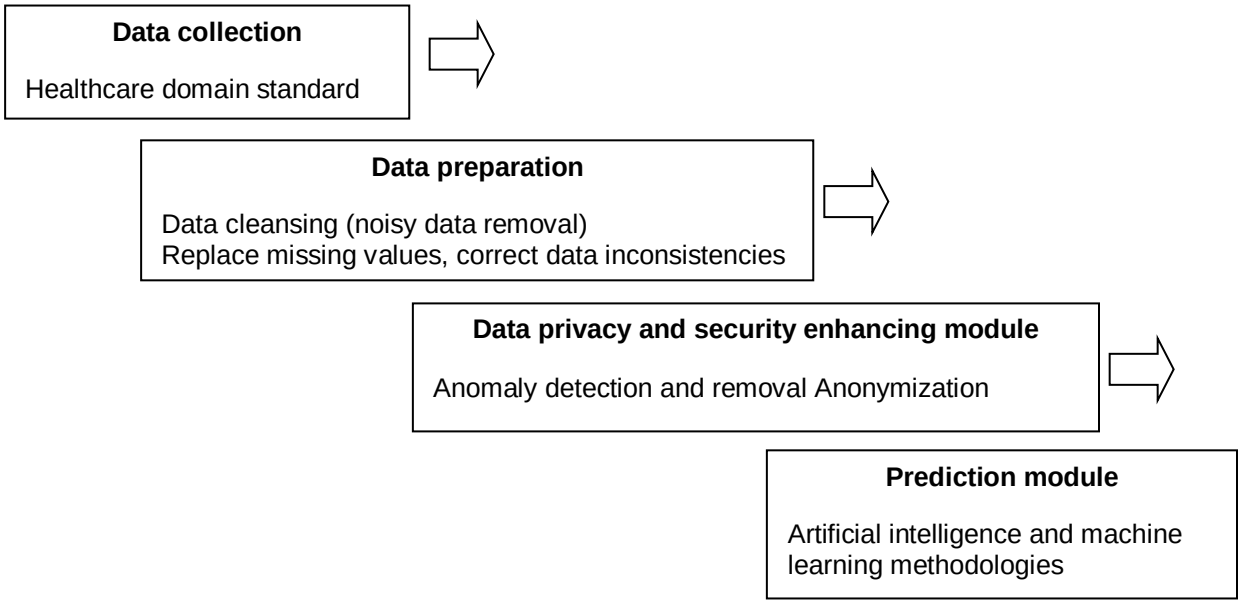
- Study currently available big data analytics techniques in healthcare domain
- Explore challenges in big data analytics considering ten Vs (volume, value, velocity, veracity, viscosity, variability, volatility, viability, validity and variety) of big data mainly related to healthcare domain
- Analyze currently available anomaly detection algorithms and their gaps to propose new methodologies which are more precise, accurate and highly efficient to ensure enriched healthcare data privacy & security
- Analyze different anonymization models/ techniques and analyze their limitations to propose improvements and novel approach to ensure enriched healthcare data privacy & security
- Evaluate recently introduced privacy and security preserving methodologies being followed mainly for healthcare domain and propose enhancements in the same by addressing their limitations to propose efficient predictive models for transforming healthcare domain

## 3.1 Expected Outcomes

- Proposed research work will provide a systematic approach to first analyze available privacy and security preserving models/ techniques/ methodologies.
- It will aim to address and overcome the challenges and shortcomings in present solutions, which will help propose novel approaches which would be precise, more accurate and efficient privacy and security preserving solutions.
- Proposed research work is focused on solving various big data analytics challenges currently faced by healthcare domain key stakeholders thus transforming future of healthcare domain

## 4. Research methodology

Proposed research work will follow below suggested research methodology. Qualitative samples (healthcare domain standard available datasets) and quantitative analysis methods will be used as a part of planned research work.


Data collection
Healthcare domain standard
Data preparation
Data cleansing (noisy data removal)
Replace missing values, correct data inconsistencies
Data privacy and security enhancing module
Anomaly detection and removal Anonymization
Prediction module
Artificial intelligence and machine learning methodologies

In our study we will use either available standard healthcare input datasets or recorded datasets from real use-cases or will generate our own datasets.

- **Input Dataset (Healthcare domain):** Dataset being considered here for planned research work will have different forms. Standard available healthcare datasets would be considered for initial work.
- **Data preparation module:** Usually, instances of healthcare datasets may have many healthcare characteristics and associated information that can be used to create a model for anomaly detection. Real-world datasets may contain noisy instances as well as an enormous number of attributes, null values, or missing values, inconsistent data because of their origin from heterogeneous sources. To improve the quality of the datasets and the performance of subsequent steps, we need to apply several data preprocessing techniques to clean this noisy

data, replace missing values, correct inconsistencies in data, and eliminate redundancies - if any. All these steps would be carried out as a part of the data preparation module.

- **Anomaly detection and removal module:** This module would identify anomaly from input dataset effectively and remove the same. Once the anomalies are identified, we remove these anomalies from the given dataset to effectively build the prediction model.
- **Anonymization module:** Anonymization refers to the complete removal of an individual's identifiers as well as the generalization of any other data that can be used to establish links to the individual.Anonymization techniques taking care of Relational healthcare data, graph-based health data would be studied, further enhanced.
- **Prediction module:** Artificial intelligence and machine learning methodologies would be applied to identify changes in classification results while doing predictions on input dataset with and without anomalies.

## 5. Importance of study (applicability to society)

The growth of world population and the rising expectations for effective treatments and overall better quality of life are putting increasing pressure on the healthcare domain. Healthcare keeps being one of the most important social and economic challenges worldwide, asking for new and more advanced solutions from science and technology year on year. Big data analytics is expected to reform the healthcare domain however there are multiple challenges (enlisted in earlier sections of research proposal) which are having adverse impact on the healthcare domain.

Proposed research work will try to address healthcare big data analytics key challenges such as data privacy, security. We will also come up with novel approaches where open challenges from existing algorithms/ models/ technologies would be addressed in efficient ways using AI/ machine learning.

## 6. Proposed work plan

| # | Tasks | Timelines | Dependencies - if any | Completion criteria | Status |
|---|---|---|---|---|---|
| 1 | Literature review | August 2021 to January 2022 | None | Synopsis preparation | Completed |
| 2 | Synopsis review and revisions with guide | February and March 2022 | Dependent on task 1 | Approved synopsis for RRC | Completed |
| 3 | RRC research proposal presentation and approval | 1st April 2022 | Dependent on task 2 | Research title approval letter from RRC | Scheduled |
| 4 | Detailed study w.r.t. research proposal objectives | April 2022 to September 2022 | None | Survey paper | Planned |
| 5 | Identification of gaps in current approaches related to 1st key BDAH challenge | October 2022 to March 2023 | Dependent on task 4 | Probable solution identification w.r.t. gaps identified | Planned |
| 6 | Validation of proposed approach for 1st key BDAH challenge | April 2023 to September 2023 | Dependent on task 5 | Publication draft creation | Planned |
| 7 | Review with guide, feedback incorporation | September 2023 to October 2023 | Dependent on task 6 | Publication in peer reviewed journal (Research center approved) | Planned |
| 8 | Identification of gaps in current approaches related to 2nd key BDAH challenge | October 2023 to March 2024 | Dependent on task 7 | Probable solution identification w.r.t. gaps identified | Planned |
| 9 | Validation of proposed approach for 2nd key BDAH challenge | April 2024 to September 2024 | Dependent on task 8 | Publication draft creation | Planned |
| 10 | Review with guide, feedback incorporation | September 2024 to October 2024 | Dependent on task 9 | Publication in peer reviewed journal (Research center approved) | Planned |
| 11 | Identification of gaps in current approaches related to 3rd key BDAH challenge | October 2024 to March 2024 | Dependent on task 10 | Probable solution identification w.r.t. gaps identified | Planned |
| 12 | Validation of proposed approach for 3rd key BDAH challenge | April 2024 to September 2024 | Dependent on task 11 | Publication draft creation | Planned |
| 13 | Review with guide, feedback incorporation | September 2024 to October 2024 | Dependent on task 12 | Publication in peer reviewed journal (Research center approved) | Planned |
| 14 | Thesis first draft creation | November 2024 to March 2025 | Dependent on all above tasks | 1st Draft of thesis ready for review | Planned |
| 15 | Thesis draft review with guide, revisions | March 2025 to April 2025 | Dependent on all above tasks | Thesis ready (post guide's approval) for next steps | Planned |

# 7. References

[1] Giuseppe Aceto, Valerio Persico and Antonio Pescape“Industry 4.0 and Health: Internet of Things, Big Data, and Cloud Computing for Healthcare 4.0”In *Journal of Industrial Information Integration*(2020) https://doi.org/10.1016/j.jii.2020.100129

[2] Shuva Paul 1, MuhtasimRiffat2, Abrar Yasir 2, Mir Nusrat Mahim 2, Bushra Yasmin Sharnali 2, Intisar TahmidNaheen 2, Akhlaqur Rahman 3 and Ambarish Kulkarni 4 “Industry 4.0 Applications for Medical/Healthcare Services”In *Journal of Sensor and Actuator Networks*(2021) https://doi.org/10.3390/jsan10030043

[3] Siva Sankara Reddy Donthi Reddy, Udaya Kumar Ramanadham“Big Data Analytics for Healthcare Organization, BDA Process, Benefits and Challenges of BDA” In *A Review Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 4, 189-196*(2017) DOI: https://doi.org/10.25046/aj020425

[4] Fatima Lalmi and LaadjalAdala“Big Data for Healthcare: Opportunities and Challenges Studies in Computational Intelligence” In *Volume 935 The Fourth Industrial Revolution: Implementation of Artificial Intelligence for Growing Business Success*(2021) DOI: https://doi.org/10.1007/978-3-030-62796-6_12

[5] Rony Chowdhury Ripan· Iqbal H. Sarker· Syed Md. “A Data Driven Heart Disease Prediction Model Through K Means Clustering Based Anomaly Detection” In *SN Computer Science (*2021) https://doi.org/10.1007/s42979-021-00518-7

[6] Krishnan Naidoo and VukosiMarivate“Unsupervised Anomaly Detection of Healthcare Providers Using Generative Adversarial Networks” In*International Federation for Information Processing 2020*(2020) https://doi.org/10.1007/978-3-030-44999-5_35

[7] Menaa Nawaz and Jameel Ahmad “Anomaly Detection and Classification of Physiological Signals in IoT- Based Healthcare Framework” In *Research Square* (2021) https://doi.org/10.21203/rs.3.rs-540398/v1

[8] Iyiola E. Olatunji, Jens Rauch, Matthias Katzensteiner and Megha Khosla “A Review of Anonymization for Healthcare Data” In*Oxford University Press*(2021) https://arxiv.org/abs/2104.06523

[9] Jayapradha. J and Prakash. M “An efficient privacy-preserving data publishing in health care records with multiple sensitive attributes: In*Proceedings of the Sixth International Conference on Inventive (2020) Computation Technologies* DOI: 10.1109/ICICT50816.2021.9358639

[10] Veena Gadad, Sowmyrani C N and Ramakanth Kumar P “An Effective Algorithm For MSA to preserve Data Privacy”In *7th International Conference Proceedings on Advanced Computing and Communication Systems (ICACCS)*(2021) ICACCS DOI: 10.1109/ICACCS51430.2021.9442019

[11] JahoonKoo,Giluk Kang and Young-Gab Kim “Security and Privacy in Big Data Life Cycle: A Survey and Open Challenges” In *MDPI Sustainability Journal* (2020) https://doi.org/10.3390/su122410571

[12] Arshia Rehman, SaeedaNaz,Imran Razzak “Leveraging Big Data Analytics in Healthcare Enhancement: Trends, Challenges and Opportunities” In *Multimedia Systems Springers Journal* (2021) https://doi.org/10.48550/arXiv.2004.09010

**Research Scholar**
Chetan Ganpat Malavade

**Research Guide**
Dr. Megha Venkatesh Jonnalagedda
Associate Professor, Department of Information Technology
Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded

**Date:** 25th March 2022.
**Place:** Nanded, Maharashtra, India